QUAL A FUNÇÃO DOS *CORPORA* NA DESCRIÇÃO DO LÉXICO?

Maria Clara Cunha
Instituto Superior de Contabilidade e Administração do Porto
Portugal
mcastro@iscap.ipp.pt

Resumo

Neste artigo procuramos reflectir sobre a função dos *corpora* na observação e análise de fenómenos de uma língua natural bem como na criação de novos recursos de exploração linguísticos que as tecnologias de informação têm vindo a potenciar e a tornar mais eficaz.

Abstract

This paper aims at reflecting upon the function of corpora in terms of the observation and analysis of linguistic phenomena that they allow as well as the design of new resources/tools that they foster and which are made available by leading new IT solutions which in turn make them increasingly more efficient and optimized.

Palavras-chave: *corpora* – léxico – fenómenos linguísticos – novos recursos/instrumentos

Key words: corpora – lexicon – linguistic phenomena – new resources/tools

É com as palavras que fazemos e descrevemos a história do mundo, a ciência e a natureza. Conceptualizamos o que conhecemos, o que nos rodeia e o que experimentamos.

O acto linguístico constitui uma referência a esse mundo que percepcionamos cognitivamente e que ordenamos intelectualmente e molda-se em diferentes soluções formais nos signos que integram as línguas naturais, pelo que todo o produto verbal configura, assim, uma mensagem conforme ao "estado de coisas" desenhado pelo saber acerca do mundo partilhado pelos falantes.

O léxico de uma língua é, por consequência, a expressão do conhecimento sobre o mundo. Este conhecimento representado lexicalmente liga-se a outras projecções - permite-nos, nomeadamente, vislumbrar uma multiplicidade de valores contidos (e, por vezes, escondidos) nas palavras.

É esta combinação de conhecimentos e sentimentos armazenados nas palavras que um falante nativo transporta no seu saber linguístico. É no léxico que uma dada comunidade linguística vaza o seu contacto e conhecimento do mundo, procurando torná-lo estável e codificado bem como um ponto de referência para outros saberes. Todavia, o léxico não é uma soma de nomenclaturas que etiquetam a realidade: a transitoriedade das coisas e do mundo, a história e o devir aninham-se no seu interior. As palavras adequam-se a cada (nova) situação, mesmo as mais intangíveis: as palavras, muito além de plasmarem conteúdos, constituem autênticos programas de representação do social ao cultural.

Então, qual a função dos *corpora*?

Actualmente, parece unânime a noção de *corpora* como fonte matricial que disponibiliza, em formato electrónico, inventariação autêntica e fidedigna das unidades de uma língua, documentadas e contextualizadas (neologismos, estrangeirismos, empréstimos, inovações lexicais e sintácticas, expressões idiomáticas, fraseologias, formantes, etc.) a partir da linguagem real que é

encontrada em diversos *corpus* de textos produzidos com intuitos diferentes e em contextos comunicativos vários (cf. Lino,1995:68): textos de língua corrente, literários, técnico-científicos, jornalísticos e didácticos [de vulgarização e/ou banalização] e que podem ser manipulados e hierarquizados *ad hoc*.

De qualquer modo, a sua definição não está isenta de alguma opacidade, como se pode verificar pelas palavras de Rastier (2004):

> *Cependant, un corpus n'est pas plus un sac de mots qu'un nébuleux intertexte. Il est structuré d'une part en fonction d'une typologie des textes, qui se réflète dans leur codage, et d'autre part, dans chaque utilisation, par des sélections raisonnées de sous-corpus.*

.

Os *corpora* apresentam diversas virtualidades aos investigadores (cf. Lino, 1995:69-70), nomeadamente permitem a observação sistemática e a análise detida de fenómenos linguísticos a diferentes níveis da descrição linguística (morfossintáctico, sintáctico, semântico e suas correlações) a partir dos quais se torna exequível, no plano das unidades lexicais e das suas relações:

- Apreciar evoluções de frequência;
- Controlar taxas e padrões de ocorrência (colocações por exemplo);
- Demonstrar fenómenos de produtividade;
- Detectar invariâncias e regularidades;
- Verificar ocorrências em contextos (hiperonímia, hiponímia, meronímia, etc.);
- Apurar a estabilização de polissemias, sinónimos e antónimos;
- Aferir a estabilização da definição de um termo.

Facultam, igualmente, a extracção de dados para a construção de novos dicionários/outros instrumentos de consulta da língua ou a sua actualização, designadamente:

- A análise de concordâncias e contextos;
- A identificação de neologismos e neónimos;
- A avaliação da incorporação de empréstimos;
- O reconhecimento de variantes terminológicas ou fraseológicas.

Recentemente, o enfoque centra-se no uso de *corpora* para extracção de candidatos a termos e sua posterior sistematização em estruturas de conhecimento informais e formais, como classificações, taxonomias e ontologias.

É importante ter *corpora* abertos, ou seja, em expansão e heterogéneos, capazes de dar conta do real estado da língua comum no seu espectro mais lato e a partir dos quais possam ser extraídos *sub-corpora* relativos a domínios específicos ou *sub-corpora* que visem acomodar análises exaustivas subordinadas ao recorte de uma área temática, com efeito como afirma Condamines (2007) *«L'utilisation des corpus ce veut um moyen d'accéder aux connaissances d'un domaine en complément ou à la place de l'expertise humaine»*.

Uma análise meticulosa e atenta dos dados resultantes de *corpora* pode formar uma base consistente para projectos práticos de recolha, construção e validação do léxico, tenha este a feição de um dicionário clássico, de um glossário terminológico ou de uma base de dados electrónica. Não obstante, os resultados gerados não devem ser usados de modo informe. A sua funcionalidade e aplicabilidade devem ser equacionadas sob o patrocínio de uma determinada moldura teórica que irá lapidar e dispor os dados, e que presidirá à própria concepção da(s) ferramenta(s) e dos seus interfaces/acessos. Como lembra Calzolari (1995:93-94):

> *Os dados provenientes de corpora não podem, obviamente, ser utilizados de modo simplista. Para se poderem tornar úteis os dados devem ser analisados à luz de uma dada hipótese teórica, na base da qual se modelará e estruturará o que seria, de outro modo, um conjunto não estruturado de dados. A melhor combinação da abordagem empírica com a abordagem teórica é aquela em que a própria hipótese teórica surge de e é guiada por sucessivas análises dos dados, e onde, por sua vez, os dados são ciclicamente aperfeiçoados e ajustados à evidência textual.*

As potencialidades dos *corpora* não se esvaziam no que até aqui nomeámos, a criação de recursos de exploração linguísticos e/ou didácticos, tais como correctores ortográficos, lematizadores, alinhadores, segmentadores, extractores, concordanceiros e glossários para a tradução automática também são importantes.

Nas últimas décadas, tem sido valorizada a importância dos *corpora* como repositórios da língua falada e escrita e como plataformas de experimentação de hipóteses linguísticas através de *softwares* de processamento da linguagem natural crescentemente optimizados, conforme nos confirma Rastier (2204) *«Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.».*

O incremento das tecnologias de informação tem viabilizado, assim, o armazenamento, tratamento e codificação de vastos *corpora* cuja exploração facilita a investigação de fenómenos linguísticos empiricamente observáveis, de uma forma cada vez mais eficaz, fiável e completa.

Os critérios de selecção dos tipos de texto que vão compor os *corpora* devem ser precisos e distintos conforme o tipo de investigação, a sua finalidade e âmbito de estudo para que seja possível coligir e seleccionar apenas os dados essenciais com vista a formar a amostra necessária. A compilação de *corpora* convoca ainda

certos princípios enformadores que apontam para a sua representatividade, exaustividade, homogeneidade e adequação sempre associadas à pertinência e utilidade do estudo a empreender. Assim, a diversidade de escolhas é grande, quer se pretendam *corpora* escritos ou orais, tematicamente paralelos ou só paralelos, monolingues, bilingues ou plurilingues; que sustentem estudos diacrónicos, diatópicos, diastráticos, diafásicos, terminológicos; etc.. Em todo o caso, será de excluir toda a variável que potencialmente possa introduzir desvios à norma ou idiossincrasias capaz de favorecer um volume indesejável de *hapax legomena*.

Os elementos obtidos permitem apontar para generalizações fundamentais para o estudo de uma língua, para a caracterização de traços linguísticos proeminentes e peculiares desta, quer ainda à captação de outra informação – a que é proporcionada por elementos que demonstram preferências ou intuições dos falantes nativos, sem prejuízo para a gramaticalidade das frases, ou a evidência da acomodação de novos termos e conceitos

Nos dias de hoje, em que o conhecimento é cada vez mais compartimentado e especializado, em que crescem as exigências de um mercado cada vez mais feroz, em que a linguagem apressada dos *media* nos esmaga diariamente; perante o tráfego contínuo de novos termos contemporâneos que nos submerge e a avidez do homem em ser actual e actualizado, emergem necessariamente outras formas e meios para comunicar, que se pretendem esquivar dos inelutáveis processos de arcaização e desfasamento da informação, daí a crescente procura e importância das publicações *online* que se tornam cada vez mais versáteis e de fácil utilização e o desenvolvimento crescente de sistemas colaborativos de imensas potencialidades bem como de novas abordagens e rumos de reflexão e de investigação científica.

BIBLIOGRAFIA:

CALZOLARI, N. (1995) «**Observação e Generalização: Análise Linguística de Verbos Declarativos Italianos com Base em Corpora Linguísticos», Actas do XI Encontro Nacional da Associação Portuguesa de Linguística (vol. I), Lisboa: APL, pp. 93-100**.

CONDAMINES, A. & AUSSENAC-GILLES, N. (2007) «Corpus et terminologie», R.T. Pédauque (ed.): *La redocumentarisation du monde »*. Toulouse: Cepadues Editions, pp.131-147 [em linha] URL: http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=acondami&subURL=index.html (consultado a 19 de Fevereiro de 2010).

LINO, Maria Teresa Rijo da Fonseca (1995) **«Da Constituição de Corpora à Lexicografia Informatizada de Especialidade», Actas do XI Encontro Nacional da Associação Portuguesa de Linguística (vol. II), Lisboa: APL, pp.** 67-92.

RASTIER, F. (2004) «Enjeux épistémologiques de la linguistique de corpus», *Texto!* [em linha], URL: http://www.revue-texto.net/index.php?id=543 (consultado a 19 de Fevereiro de 2010).

SINCLAIR, J. (1995) **«Tipologia Textual EAGLES», Actas do XI Encontro Nacional da Associação Portuguesa de Linguística (vol. I), Lisboa: APL, pp.** 39-72.